BIOMETRICS 59, 1184–1185
December 2003

On Use of Bivariate Survival Models with Cure Fraction

Nilanjan Chatterjee

Division of Cancer Epidemiology and Genetics, National Cancer Institute, 6120 Executive Blvd, EPS 8038, Rockville, Maryland, U.S.A.
email: chattern@mail.nih.gov

and

Joanna H. Shih

Division of Cancer Treatment and Diagnosis, Biometric Research Branch, National Cancer Institute, 6130 Executive Blvd, EPN 8132, Rockville, Maryland, U.S.A.
email: jshih@mail.nih.gov

In the current issue of the journal, Wienke et al. have considered bivariate (or multivariate) survival modeling in presence of cure fraction, a problem that was motivated in the article *"A Bivariate for Modeling Cure-Mixture Approach Familial Association in Diseases"* (Chatterjee and Shih, 2001). In the discussion, the authors stated that their work extended our original approach in two ways:

- Use of the more flexible correlated gamma-frailty model, and
- direct estimation based on maximum-likelihood method instead of our proposed two-stage estimation.

Moreover, the authors argued that the application of the methodology to their twin data set was more appropriate as an illustrative example than the Washington Ashkenazi Study family data we used. In what follows, we will first comment on the above specific issues.

Correlated Gamma Frailty Model

In our original article, we proposed the methodology in the very general setting of copula model (Genest and Mackay, 1986; Marshall and Olkin, 1988). We illustrated the application of the methodology using three specific distribution of the general class, namely Clayton's (1978), Frank's (1979), and Stable family (Hougaard, 1986). It is true that all of these three models correspond to an underlying shared frailty model. The class of Copula models, however, are much more general and can include correlated frailty models. In fact, the correlated frailty model the authors proposed to use as an extension of our methodology is a special case of the copula model with the copula function given by

$$C(u,v) = u^{1-\rho}v^{1-\rho}\left(u^{-\sigma^2} + v^{-\sigma^2} - 1\right)^{-\frac{\rho}{\sigma^2}} \qquad (1)$$

Thus, all the inferential methods we proposed in our original article will be also applicable to this correlated frailty model.

Two-Step vs. One-Step Estimation Method

We recommended two-stage estimation, not because we were not able to do one-step MLE for the parametric model, but for the sake of robustness and flexibility. The quasi-likelihood we defined as L_2 in equation (3) of Chatterjee and Shih (2001) is, in fact, the actual likelihood for bivariate data. In a completely parametric problem, it is straightforward to maximize this likelihood to obtain the parameter estimates. We, however, based on the reasons described below, would like to consider use of semiparametric model that allows completely unspecified or nonparametric marginal survivor functions.

In general, in analysis of clustered data when one wants to fit and compare different parametric correlation models, it is often recommended that the mean structure of the data (corresponding to marginal distributions) should be modelled as elaborately as possible (see, e.g., Diggle, Liang, and Zeger, 1994, Section 5.3.1)—if possible nonparametrically, so that the correlation is not confounded with underspecification of the mean model. In our two-stage estimation method, we can easily estimate the marginal distributions nonparametrically at stage one. At the second stage, we assess and compare adequateness of different parametric correlation models with the fixed marginal distribution estimated from stage one. Thus, the two-step procedure gives a fairly simple method for fitting the appealing semiparametric model that allows nonparametric marginal distributions and parametric correlation structure.

In the context of cure model, it may be even more important to consider a nonparametric approach for estimation of the marginal distribution. We note that cure modelling is essentially a latent variable approach and that whether a subject is "cured" or not is never directly observable, due to censoring. If, in truth, the hypothesized "cured" population does not exist, the cure model faces lack of identifiability and interpretation. Thus, before considering a cure modelling approach, some exploration of the data is needed to test for evidence in the data that would suggest presence of

"cured" population. Maller and Zhou (1992, 1995) suggested an informal and formal method for testing for presence of "cured" individuals, based on the degree of the leveling of the nonparametric Kaplan-Meier incidence curve. Informally speaking, the basic idea behind this approach is that existence of sufficient follow-up without event, after the last event observed in the data, is an indication of presence of "cured" subjects in the data. In this case, the value at which the Kaplan-Meier estimator levels off after the last event can be taken as a reliable measure of the cure fraction. In a parametric approach, although allowing for a cure parameter could give a better fit of the data with an estimate of ϕ (cure fraction) $>$ 0, one may not be able to distinguish the situation whether the better fit is due to the fact that there is a truly "cured" population or is it because of the lack of fit of the parametric survival model. Once the presence of cured population has been established based on a nonparametric approach, if suitable, one can consider a parametric specification for the marginal survivor function for the susceptible (not cured) subjects. In the application involving the Washington Ashkenazi Study in Chatterjee and Shih (2001), for example, we found a Weibull model gave a very good fit to the marginal risk of breast cancer among the susceptible subjects in this data (see Chatterjee and Shih, 2001, Figure 1).

Data Application

The authors considered an application of the proposed methodology using breast cancer data from Swedish twin registry data of twins. We feel the authors have not made full use of this rich data set. First, a nonparametric Kaplan-Meier curve (similar to Figure 1 in Chatterjee and Shih, 2001) for breast cancer incidence in this data would have been very informative. As we discussed above, such a curve could give readers a much better sense of whether the cure modelling approach is appropriate for this data. In our analysis of the Washington Ashkenazi data, we found that the oldest incidence had occurred at age 91, after which the Kaplan-Meier curve leveled off. Between age 91 and the oldest follow-up age of 103, there were 178 subjects, with an average follow-up of about age 94.3 years, who experienced no breast cancer event. Presence of such a significant number of subjects who were followed until old age without any breast cancer event gave us some indication of "insusceptible" individuals in this population.

A key finding of our analysis of breast cancer data was that, after accounting for possibility of correlation between susceptibilities (cure probabilities) of pairs of relatives, we observed only a modest association between their age-at-onset of the disease. We observed that when the possibility of correlation between susceptibilities of two relatives was not allowed, the correlation between the age-at-onset of the relatives seemed much stronger, as it absorbed the correlation between the susceptibilities. In light of our finding, it would have been interesting if the authors had reported the estimate and the confidence interval of the pairwise odds ratio between the susceptibility status of two relatives, defined as γ in Chatterjee and Shih (2001). The authors only noted that the model that allowed for such correlation did not significantly fit better than the model that assumed independence.

As the authors did not interpret the results of their application (Chatterjee and Shih, 2001, Table 1), a number of issues remain unclear. First, what do the authors gain by considering the more complex two-parameter correlated frailty model? Does this model give a better fit than the one-parameter shared frailty model we considered? When the possibility of cure fractions is allowed, why is the correlation between monozygotic twins exactly 1.0? Is this a constraint? What is the heuristic explanation that the estimate of the parameter ρ is so sensitive to whether a cure fraction is allowed in the model or not?

Concluding Remarks

In summary, our original article (Chatterjee and Shih, 2001) and the current article by Wienke et al. both demonstrate that, for analyzing correlation in the survival analysis framework, it can be important to account for presence of "cured/insusceptible/immune" subjects, if such a population truly exists. The presence of such a population cannot be directly tested, as cure status of subjects is unobservable, due to censoring. For identifiability and interpretation of parameters, however, it is important to first examine if data provide some evidence about presence of a "cured" population. In a parametric survival analysis setting, superior fit of a model that allows for cure parameters may not necessarily indicate presence of "cured" subjects. Nonparametric methods, such as those suggested by Maller and Zhou (1992, 1995), should be used to examine if there is empirical evidence in the data of "cured" subjects. If the data support evidence of cured subjects, then a cure modelling approach can provide important insight in the context of both univariate and multivariate survival analysis.

References

Chatterjee, N. and Shih, J. (2001). A bivariate cure-mixture approach for modeling familial association in diseases. *Biometrics* **57,** 779–786.

Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* **65,** 141–151.

Diggle, P. J., Liang, K. Y., and Zeger, S. L. (1994). Analysis of longitudinal data. Oxford: Oxford University Press.

Frank, M. J. (1979). On the simultaneous associativity of $F(x, y)$ and $x + y - F(x, y)$. *Aequationes Mathematicae* **19,** 194–226.

Genest, C. and Mackay, J. (1986). The joy of copulas—Bivariate distribution with uniform marginals. *American Statistician* **40,** 280–283.

Hougaard, P. (1986). A class of multivariate failure time distributions. *Biometrika* **73,** 671–678.

Maller, R. A. and Zhou, S. (1992). Estimating the proportion of immunes in a censored sample. *Biometrika* **79,** 731–739.

Maller, R. A. and Zhou, S. (1995). Testing for the presence of immune or cured individuals in censored survival data. *Biometrics* **51,** 1197–1205.

Marshall, A. W. and Olkin, I. (1988). Families of multivariate distributions. *Journal of the American Statistical Association* **83,** 834–841.

Received June 2003. *Revised n.a.*
Accepted June 2003.